Genomic regions exhibiting positive selection identified from dense genotype data

Christopher S. Carlson,[1,3] Daryl J. Thomas,[2] Michael A. Eberle,[1] Johanna E. Swanson,[1] Robert J. Livingston,[1] Mark J. Rieder,[1] and Deborah A. Nickerson[1]

[1]*Department of Genome Sciences, University of Washington, Seattle, Washington 98195-7730, USA;* [2]*Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064-1099, USA*

The allele frequency spectrum of polymorphisms in DNA sequences can be used to test for signatures of natural selection that depart from the expected frequency spectrum under the neutral theory. We observed a significant ($P = 0.001$) correlation between the Tajima's D test statistic in full resequencing data and Tajima's D in a dense, genome-wide data set of genotyped polymorphisms for a set of 179 genes. Based on this, we used a sliding window analysis of Tajima's D across the human genome to identify regions putatively subject to strong, recent, selective sweeps. This survey identified seven Contiguous Regions of Tajima's D Reduction (CRTRs) in an African-descent population (AD), 23 in a European-descent population (ED), and 29 in a Chinese-descent population (XD). Only four CRTRs overlapped between populations: three between ED and XD and one between AD and ED. Full resequencing of eight genes within six CRTRs demonstrated frequency spectra inconsistent with neutral expectations for at least one gene within each CRTR. Identification of the functional polymorphism (and/or haplotype) responsible for the selective sweeps within each CRTR may provide interesting insights into the strongest selective pressures experienced by the human genome over recent evolutionary history.

[Supplemental material is available online at www.genome.org.]

According to the theory of neutral molecular evolution (Kimura 1983), the vast majority of DNA sequence-level polymorphism is selectively neutral in a population, and therefore, most of the observed diversity represents a balance between the introduction of new polymorphism by mutation and the extinction of existing polymorphism by genetic drift. Under the mutation/drift model as well as appropriate demographic assumptions regarding population size, random mating, recombination, and mutation rate, it is possible to predict the expected site frequency spectrum (SFS) of a region (Watterson 1975; Ewens 1979).

A number of statistical tests have been devised that compare an observed SFS against neutral theory predictions. One of the most frequently used tests is Tajima's D (Tajima 1989), a comparison of nucleotide diversity estimated from the number of polymorphic sites observed in a given set of chromosomes against nucleotide diversity estimated from the allele frequency of the polymorphic sites. Other tests of SFS against neutral expectations exist, including Fu and Li's D (Fu and Li 1993), which is based upon the number of singleton derived alleles observed, and Fay and Wu's H (Fay and Wu 2000), which is based upon the frequency spectrum of nonancestral alleles. By using these tests, a substantial number of genes have been identified where the observed SFS is inconsistent with neutrality, including *ABO* blood group (Seltsam et al. 2003), the major histocompatibility antigens (*HLA*) (Hughes and Yeager 1998), *lactase* (Bersaglieri et al. 2004), and *TRPV6* (Akey et al. 2004; Stajich and Hahn 2005). Genes with an excess of high-frequency variation (observed as a positive Tajima's D, e.g., *ABO*, *HLA*) are consistent with balancing selection, while genes with an excess of low-frequency variation (significantly negative Tajima's D, e.g., *lactase*, *TRPV6*) are consistent with positive selective pressure, where an advantageous variant (haplotype) has recently replaced most of the variation in a region. Genes subjected to a recent selective sweep, where the advantageous allele has not yet become the major allele (e.g., *Duffy* [Hamblin and Di Rienzo 2000; Hamblin et al. 2002] and *CCR5* [Libert et al. 1998]) do not have easily detectable departures using Tajima's D, but they may be detected by using Fay and Wu's test. Genes under recent balancing selection (e.g., *G6PD* [Verrelli et al. 2002], β-*globin* [Straus and Taylor 1981]) are more difficult to detect and may not be detectable by any standard nucleotide diversity test. Other simple tests of neutrality exist such as the Ewens-Watterson haplotype diversity test (Ewens 1972), haplotype lineage diversity tests (Verrelli et al. 2002), and haplotype extent tests (Sabeti et al. 2002). In addition, tests for geographically restricted selective pressure measured by divergence between populations are also available (F_{st}) (Weir and Cockerham 1984). Several recent studies have evaluated the utility of combining multiple tests to identify genes showing signatures of natural selection and have introduced more sophisticated approaches to evaluate the robustness of these analyses to the neutral model. It is noteworthy that in these studies, all genes with significantly negative Tajima's D under the simplest neutral models were robust to a range of demographic parameters (Akey et al. 2004; Stajich and Hahn 2005), although questions about some of the tested models have been raised (Thornton 2005).

The identification of >10 million single nucleotide polymorphisms (SNPs) across the human genome, and the emergence of large-scale data sets of genotypes at a subset of these SNPs in multiple populations are leading to many insights into the patterns of sequence variation across the human genome (Sabeti et al. 2002; Clark et al. 2003a,b; Nielsen et al. 2005; Williamson et al. 2005). Here we describe the application of Tajima's D to a

[3]Corresponding author.
E-mail csc47@u.washington.edu; fax (206) 221-6498.
E-mail debnick@u.washington.edu; fax (206) 221-6498.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.4326505. Freely available online through the *Genome Research* Immediate Open Access option.

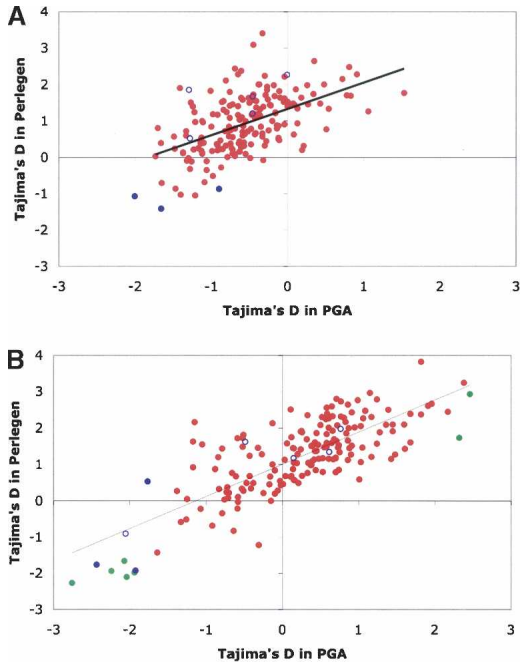

Figure 1. Comparison of Tajima's D between Perlegen and SeattleSNPs data sets. For each gene, Tajima's D was calculated from complete resequencing data in the SeattleSNPs data set, or from the region spanning 10 kb upstream of the transcript, the full transcript, and 10 kb downstream of the transcript in the Perlegen data. (*A*) Tajima's D from Perlegen vs. Tajima's D from SeattleSNPs for AD population. (*B*) Tajima's D from Perlegen vs. Tajima's D from SeattleSNPs for ED population. Genes previously resequenced by SeattleSNPs are shown in red, with a trend line representing a linear regression on the data. Genes resequenced as part of the present study are shown as purple dots, with filled circles indicating that the gene lay within a CRTR in the population being plotted. The seven SeattleSNPs genes with robust signatures of selection in SeattleSNPs data are shown in green (Akey et al. 2004).

dense genotype data set (the Perlegen data set) (Hinds et al. 2005). We compared Tajima's D in completely resequenced human genes (SeattleSNPs, http://pga.gs.washington.edu) with the Perlegen data set and observed strong correlations in both African American and European American data for coding and noncoding regions. Based on these correlations, we used an empirically derived sliding window distribution of Tajima's D to examine the autosomal regions of the human genome. We identified large regions of the genome with an excess of rare variation in each of three populations, consistent with strong and recent selective sweeps that would lead to rapid increases in the frequency of an advantageous polymorphism and/or haplotype within each region (Smith and Haigh 1974). To confirm the validity of this approach, eight genes (defined as known genes or expressed sequence tags [ESTs]) from six such regions were selected for directed resequencing to obtain the full SFS and directly assess Tajima's D. Our results are consistent with a strong, recent, positive selective pressure in each of these resequenced regions, and they demonstrate the utility of dense genotype data in identifying such regions across the genome.

Results

To evaluate the utility of the Perlegen data in evaluating Tajima's D, we compared genes resequenced by SeattleSNPs in the same African-descent (AD) and European-descent (ED) individuals. Tajima's D was compared for all autosomal genes where at least five SNPs within 10 kb of the transcript were polymorphic in the Perlegen data set. The final data set consisted of 179 genes meeting this criterion in at least one population, with 178 genes in the AD population and 173 genes in the ED population. The mean value for Tajima's D in the Perlegen data (0.94 for AD, 1.25 for ED) was substantially higher than in the SeattleSNPs data (−0.54 for AD, 0.26 for ED), as expected given an ascertainment bias toward high-frequency SNPs in the Perlegen data. A significant correlation was observed for Tajima's D between these data sets in both populations (Fig. 1A,B for AD and ED, respectively). The correlation between these two data sets was stronger in the ED population ($R^2 = 0.59$) compared with the AD population ($R^2 = 0.28$) but was significant in both populations ($P = 0.001$ by Student's *t*-test). The observed correlation is based on a comparison of genic regions, and it is unclear how well the correlation extrapolates to large intergenic regions, as selective pressures (and therefore site frequency spectra) could be different in such regions. However, the SeattleSNPs data consist of 6% coding sequence, 4% UTR sequence, 70% intronic sequence, and 20% flanking intergenic sequence. In the SeattleSNPs data, no significant differences were observed in diversity between intronic sequence and the flanking intergenic sequence, so the observed correlation applies at least to proximal intergenic sequences.

The correlation between these data sets suggested that the Perlegen data can be used to survey the genome for regions exhibiting extreme values of Tajima's D, so we applied a sliding window analysis to all three populations genotyped by Perlegen. The distribution of Tajima's D values for a 100-kbp sliding window is shown in Figure 2. The average windowed Tajima's D in the AD population was 1.20 (±0.73 SD) with a range of −2.41 to 4.03. In the ED population, the average Tajima's D was 1.40 (±1.01 SD) with a range of −2.80 to 4.34, while in the Chinese (XD) population the average was 1.45 (±1.13 SD) across a range of −3.11 to 4.42. These distributions were substantially skewed in all three populations, with a heavier tail to the distribution at low values.

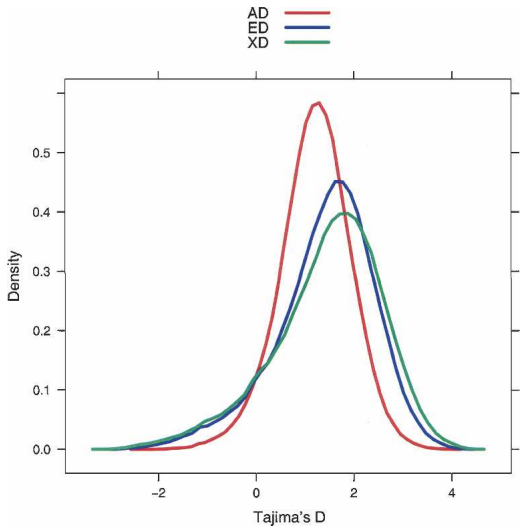

Figure 2. A probability density plot of the distribution of Tajima's D in the sliding windows is shown for each population. All three distributions depart significantly from a normal distribution, most noticeably in the heavy tail at low values in each population.

Results from the sliding window analysis are depicted across a 50-megabase segment of chromosome 1 (chr1, 1–50,000,000) in Figure 3, and results were similar across the rest of the genome. Tracks displaying the sliding window data for each chromosome are available on the UCSC Genome Browser (Kent et al. 2002). To identify regions recently subject to strong selective pressure, we developed a qualitative algorithm for identification of Contiguous Regions of Tajima's D Reduction (CRTRs) in the windowed data, as described in the Methods, and applied this independently to each population. The CRTRs identified by this approach are listed in Table 1: Seven CRTRs were identified in the AD population, 23 in the ED population, and 29 in the XD population. Four CRTRs overlapped between populations: chr20, 20360000–20690000, overlapped between the AD and ED populations, while chr11 (37980000–38290000), chr16 (46050000–46340000), and chr18 (28640000–29150000) overlapped between XD and ED populations. The majority of CRTRs spanned 300,000–400,000 bp (20–30 windows), although several large CRTRs spanned more than half a megabase in either the ED or the XD population. For example, nearly a megabase of chromosome 1 near the *CLSPN* gene (chr1, 35220000–36210000) was observed as a single CRTR in the ED population, and CRTRs spanning >600 kb were observed on chromosomes 1 (chr1, 92220000–93030000) and 2 (chr2, 108350000–109120000) in the XD population.

A major assumption in the interpretation of SFS is that ascertainment of the SNPs genotyped is unbiased or at least consistently biased such that the positive shift in Tajima's D is similar across the genome. The SNPs genotyped in the Perlegen map were drawn from three sources: Perlegen's internal resequencing project (class A), dbSNP validated SNPs (class B), and dbSNP unvalidated SNPs (class C). Because these three classes show different SFS, with classes A and C enriched for rare variants relative to class B, each class would be expected to show a different bias in the SFS. In theory, it is possible to correct an observed SFS for the ascertainment scheme used to select SNPs (Nielsen et al. 2004), but in this case because SFS has been assessed within specific ethnic groups, it would be inappropriate to use this correction because the SNPs were ascertained in a global mixture of ethnic groups (R. Nielsen, pers. commun.). In order to investigate the possibility that a biased SFS from the three classes could account for the identified CRTRs, we examined the relative frequency of each SNP class within the CRTRs and compared it against genome-wide averages. Genome-wide, 79.9% of SNPs were class A, 18.5% were class B, and 1.6% were class C. Since SNP class data were mapped to the hg16 build of the human genome, we remapped the CRTR results from hg17 to hg16. One of 55 unique CRTRs mapped to two regions in the hg16 genome build and was excluded from further analysis. Among the 54 CRTRs uniquely mapped to build hg16, we found a modest but significant excess of class B (20.6%, $P = 0.001$ by χ^2 test) and class C SNPs (1.9%, $P = 0.006$ by χ^2 test).

To further assess whether SNP ascertainment bias might account for the CRTRs, we examined each CRTR independently (Table 1). The relative rarity of class C meant that fewer than five class C SNPs were expected in most CRTRs, so we merged classes A and C for this comparison. The proportion of class B SNPs ranges from 1.8%–85.3%, with just six CRTRs >50%. In 28 CRTRs we observed a significant departure in the frequency of class B SNPs from genome-wide averages, after Bonferroni correction for 54 tests. Tellingly, 16 of these departures were toward an excess of class B, and 12 were toward an excess of the non-B classes. Thus, although class B SNPs are modestly enriched in the CRTRs on average, only a minority of CRTRs are enriched for such SNPs, with a nearly even number of CRTRs enriched for type A and C SNPs. We interpret this as evidence that SNP class (and thus SNP ascertainment bias) has not substantially biased CRTR identification.

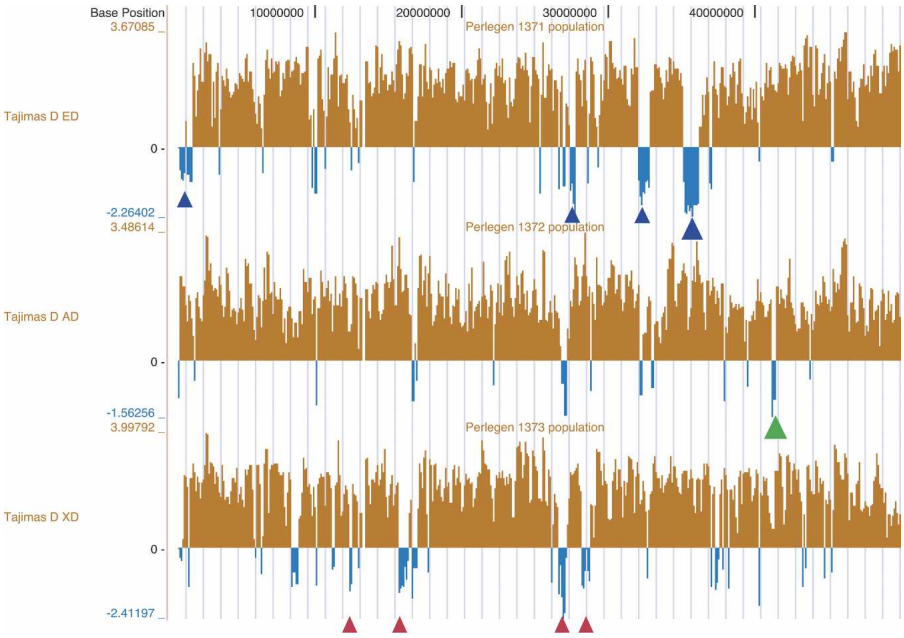

Figure 3. Tajima's D in 100-kbp sliding windows with 10-kbp steps is shown across the first 50 megabases of chromosome 1. Several CRTRs are visible, including a region near 35M in the ED population containing *CLSPN* (large blue arrowhead) and a region near 41M in the AD population spanning *CTPS*, *FLJ23878*, and *SCMH1* (large green arrowhead). CRTRs at the less stringent 5% level are also indicated in the ED population as small blue arrowheads and in the XD population as small red arrowheads.

It is also possible that low recombination rates biased our survey toward the identification of low diversity regions that coincide with low recombination rates. The average recombination

Table 1. CRTRs

Genome coordinates[a]	Windows <1% threshold[b]	Total SNPs[c]	SNP class A/B/C[d]	Monomorphic AD/ED/XD[e]	Rec rate[f]	Known genes[g]
CRTRs in the AD population						
chr1: 41080000–41470000	29/30	137	106/27/4	26+/10−/19	1.2	**CTPS, FLJ23878, SCMH1**
chr2: 95300000–95940000	45/55	70	36/28/6+	4/13/10	0.1	TRIM43, AK024144 TRIM51, OR5W2, OR10AG1, AK055955, OR5F1, OR5AS1, OR8I2, OR8H2, OR8H3, OR8J3, OR8K5, OR5J2, OR5T2, OR5T3
chr11: 55340000–55750000	32/32	431	390/39/2−	42/37−/279+	0.0	OR5T1, FLJ43276, ZSCAN2, SCAND2, FLJ14888
chr15: 82780000–83180000	24/31	—[h]	—[h]	—[h]	0.0	NMB, SPC18, ZNF592
chr16: 14460000–14760000	21/21	39	35/4/0	11+/5/34+	1.8	PARN, BFAR, PLA2G10
chr20: 20360000–20720000[j]	27/27	164	133/31/0	11/63+/81+	2.4	AB033098
chr22: 26700000–27190000	35/40	267	259/7/1−	31+/91+/115+	2.4	PITPNB, KIAA1043, SFPQ, ZNF262, PKD1L, NCDN, TFAP2E, PSMB2
CRTRs in the ED population						
chr1: 35220000–36210000	87/90	192	155/29/8	13/84+/41	0.9	**CLSPN,** EIF2C4, EIF2C1, EIF2C3
chr2: 84470000–84810000	25/25	192	101/88/3+	22/83+/75+	0.2	SUCLG1, FLJ37357
chr2: 162820000–163240000	25/33	228	221/5/2−	75+/88+/72	0.4	**GCG,** FAP, IFIH1, GCA, KCNH7
chr3: 89690000–90110000	33/33	110	47/61/2+	3/31/32	0.1	[i]
chr4: 32930000–33560000	50/54	276	215/50/11	17/92+/141+	0.6	[i]
chr4: 148880000–149280000	31/31	200	176/18/6−	11/85+/85+	0.8	TMEM34, ARHGAP10
chr6: 75060000–75360000	21/21	139	116/23/0	16/34/23	0.0	[i]
chr6: 75580000–75890000	22/22	142	70/69/3+	29+/37/48	0.0	[i]
chr6: 84500000–84800000	21/21	118	96/22/0	7/31/50+	0.2	NCB50R
chr6: 105600000–105980000	22/29	166	119/37/10	2/40/59	0.8	BVES, POPDC3, PREP, AK025690, ARPC1A, ARPC1B, PDAP1, G10, PTCD1, CPSF4, ATP5J2, ZNF394, ZFP95, VIK
chr7: 98460000–99080000	52/53	135	76/51/8+	7/54+/53	1.0	ZNF498, CYP3A5, CYP3A7, CYP3A4, VCIP135, FLJ11267, PTTG3, SGKL, FLJ25692
chr8: 67650000–68280000	47/54	93	67/15/11	15+/43+/89+	0.4	COPS5, FLJ22490
chr10: 68630000–68930000	16/21	122	103/15/4	30+/36/27	0.7	CTNNA3, P4HA1, NUDT13, HSGT1, DNAJC9, MRPS16
chr10: 74340000–75090000	63/66	177	136/38/3	14/46/32	0.0	TTC18, ANXA7, ZMYND17, PPP3CB, USP54
chr11: 37980000–38360000[k]	29/29	281	201/78/2+	12/91+/102	0.1	[i]
chr12: 42710000–43010000	20/21	174	162/9/3−	24+/55+/66	0.6	DKFZp434K2435
chr12: 87490000–87840000	26/26	124	98/24/2	9/46+/61+	0.4	[i]
chr14: 44280000–44700000	33/33	177	142/29/6	15/39/58	0.6	BTBD5, KIAA0423, PRPF39
chr16: 46030000–46340000[k]	17/22	68	56/8/4	10/13/17	0.3	PHKB
chr18: 28630000–29170000	40/45	307	232/73/2	7/86+/101	0.6	C18orf34
chr18: 65710000–66040000[k]	24/24	247	228/17/2−	11/39/214+	1.8	CD226, RTTN
chr19: 47540000–47920000	28/29	67	49/17/1	3/16/37+	0.5	LIPE, CEACAM1, CEACAM8
chr20: 20360000–20690000[j]	21/24	151	124/27/0	11/62+/79+	2.4	AB033098
CRTRs in the XD population						
chr1: 72310000–72790000	39/39	358	317/38/3−	26/184+/217+	0.2	NEGR1
chr1: 87010000–87320000	18/22	103	83/18/2	14/9/39	0.9	SEP15, HS2ST1
chr1: 92220000–93030000	68/72	252	150/96/6+	20/37/79	0.6	BTBD8, GLMN, FLJ13150, GFI1, EVI5

(Continued)

Table 1. *Continued*

Genome coordinates[a]	Windows <1% threshold[b]	Total SNPs[c]	SNP class A/B/C[d]	Monomorphic AD/ED/XD[e]	Rec rate[f]	Known genes[g]
chr1: 103130000–103460000	24/24	201	145/56/0+	12/52/90+	0.3	COL11A1
chr2: 84540000–84910000	28/28	197	133/59/5+	17/90+/65	0.2	SUCLG1, FLJ37357, DNAH6, AB051484, SULT1C2, GCC2, LIMS1, RANBP2, FLJ32745
chr2: 108350000–109120000	68/68	360	309/37/14−	17/79/195+	0.9	**EDAR**
chr2: 177390000–177730000	20/25	447	439/8/0−	24/125+/353+	1.2	[i]
chr2: 189140000–189570000	31/34	245	204/41/0	17/85+/124+	0.3	GULP1, DIRC1
chr2: 194650000–194990000	24/25	102	15/87/0+	4/88+/37	0.2	[i]
chr3: 17570000–17890000	21/23	115	81/31/3	17+/18/69+	0.6	TBC1D5
chr3: 25700000–26250000	46/46	229	154/70/5+	10/72+/107+	1.2	NGLY1
chr4: 41670000–42050000	29/29	181	150/29/2	2/20/59	1.2	TMEM33, SLC30A9, FLJ43695
chr5: 117360000–117700000	25/25	227	161/64/2+	4/34/63	0.9	[i]
chr6: 126650000–127130000	39/39	110	79/31/0	6/30/37	0.8	[i] EPHB6, TRPV6, TRPV5, KEL, OR9A2, OR6V1
chr7: 142030000–142360000	24/24	211	44/162/5+	7/91+/103+	0.3	C7orf34
chr8: 50580000–51170000	45/50	432	364/65/3	13/162+/193+	0.4	SNTG1
chr11: 37820000–38290000[k]	38/38	377	295/80/2	13/108+/144+	0.4	[i]
chr12: 22370000–22700000	24/24	134	124/8/2−	18/53+/49	0.9	KIAA0528
chr12: 86840000–87360000	41/43	147	88/44/15+	6/23/43	0.5	FLJ35821, AL137488, Cep290, SMILE
chr13: 19040000–19390000	26/26	119	102/12/5	7/32/37	0.8	HSMPP8, PSPC1, ZNF237
chr13: 62440000–62760000	23/23	210	157/53/0	16/26/47	0.0	[i]
chr14: 67310000–67770000	36/37	257	229/25/3−	5/70+/159+	0.3	RAD51L1
chr15: 61550000–62070000	43/43	262	233/26/3−	10/42/195+	0.6	USP3, FBXL22, HERC1, DAPK2, BC058863, BCL7C, CTF1, FBXL19, HSD3B7
chr16: 30700000–31070000	26/28	48	12/34/2+	2/20+/27+	0.4	STX1B2, STX4A, FLJ13479, KIAA0296, VKORC1
chr16: 45180000–45500000	23/23	58	22/31/5+	2/32+/37+	0.0	VPS35, ORC6L, LOC91807, LOC388272
chr16: 46050000–46490000[k]	28/35	105	87/14/4	9/16/24	0.3	PHKB, CBFB, Lin10, TRADD, FBXL8, HSF4, NOL, E2F4, ELMO3, LRRC29, HSPC171, FHOD1, SLC9A5
chr16: 65590000–66060000	36/38	106	85/7/14	7/36+/51+	0.3	AK097481, FLJ11004, CGI-38, ZDHHC1
chr17: 61300000–61640000	25/25	187	140/46/1	2/12−/30	1.4	AK096929
chr18: 28640000–29150000[k]	35/42	300	228/70/2	7/83+/98	0.6	C18orf34

[a]Genome coordinates in the hg17 build.
[b]Fraction of 100-kb windows <1% empiric threshold.
[c]Total number of Perlegen SNPs in the window.
[d]Perlegen class SNP counts. 106/27/4 indicates 106 class A, 27 class B, and 4 class C. + indicates significant excess of B relative to genome average, − indicates significant deficit of B relative to genome average.
[e]Count of monomorphic SNPs in each population. + indicates significant excess of monomorphic SNPs relative to genome average in a given population, − indicates significant deficit of monomorphic SNPs.
[f]Recombination rate for flanking megabase from deCODE sex averaged map.
[g]Known genes within CRTR derived from the UCSC Genome Browser Known Genes (November 22, 2004) track. Bold are resequenced genes.
[h]Not mapped in hg16.
[i]No known gene within CRTR.
[j]CRTR shared between ED and AD populations.
[k]CRTRs shared between ED and XD populations.

rate for regions flanking the CRTRs was 0.67 cM/Mb (Table 1), which is significantly lower than the genome-wide average of 1.13 cM/Mb, but only four of the regions fell into recombination deserts with an estimated recombination rate of 0.0 cM/Mb. However, the observed correlation between CRTRs and low recombination regions is also expected if the CRTRs are the product of selective pressure, because the region that is swept along with an advantageous allele will be larger in regions of low recombination. Thus, there does appear to be a modest bias in the data toward identification of CRTRs in regions with low recombination rates.

To assess how well CRTRs in the Perlegen data predict Tajima's D in resequencing data, we selected eight targets from six population-specific CRTRs for resequencing (Table 2). In each CRTR, targets were chosen for one or more of the following reasons: dramatic allele frequency differences between populations in the Perlegen data (*EDAR*, *CLSPN*), central position within the CRTR (*EDAR*, *CLSPN*, *SCMH1*, *FLJ23878*), important gene function (*GCG*), or the target was a spliced EST in a CRTR without any known genes (AW183861, BX115137). *CTPS* and *FLJ23878* were included because these genes, in combination with *SCMH1*, accounted for all of the known genes within a single CRTR in the AD population. Values of Tajima's D below -2 for resequencing data are significant under the simplest neutral models (Tajima 1989) and have been asserted to be robustly inconsistent with neutrality under a variety of demographic models (Akey et al. 2004; Stajich and Hahn 2005), although questions regarding the bottleneck models have been raised (Thornton 2005). In four of six CRTRs resequenced, at least one gene was identified with a Tajima's D below -2 in the appropriate population, and the other two CRTRs also show a substantial excess of low-frequency variation in the SFS. In AW183861 the observed Tajima's D was -1.92 in the ED population, placing it in the same range as DCN, which has previously been suggested as a target of selective pressure (Akey et al. 2004; Stajich and Hahn 2005). The other gene with marginal Tajima's D was GCG, which was selected because of its importance in glucose metabolism but lay at the boundary of the CRTR. In this case, the promoter and transcript of the gene (bases 1–11349) lie within the CRTR and showed a significant excess of rare polymorphism in the SFS (Tajima's D = -2.48), while the region 3′ of the transcript lies beyond the CRTR and showed a less extreme SFS (Tajima's D = -0.45). Thus, in all six resequenced regions we observed a strong trend toward a negative Tajima's D.

Although we demonstrate that CRTRs identified from the Perlegen data predict extreme Tajima's D in resequencing data, dramatic departures from the expected SFS might be expected to occur at random in the genome, so we investigated whether the underlying nucleotide diversity of these regions was also consistent with selective pressure. Tajima's D detects departures from the expected SFS under neutral assumptions by comparing two measures of nucleotide diversity, θ and π. The absolute values of these statistics can also provide evidence for selective pressure, as both are reduced as an advantageous allele nears fixation. Estimated from the 179 SeattleSNPs genes previously compared against the Perlegen data, the average π was 9.02 ($\pm$3.80) $\times 10^{-4}$ in AD and 7.17 ($\pm$4.00) $\times 10^{-4}$ in ED populations, and the average θ was 10.44 ($\pm$3.14) $\times 10^{-4}$ in AD and 6.48 ($\pm$2.68) $\times 10^{-4}$ in ED populations. All of the CRTR genes resequenced show trends toward reduced π in the appropriate population, and most also show reduced θ (Table 2), which is consistent with selection, although low absolute nucleotide diversity might also be attributable to reduced mutation rates in these regions.

We further examined the possibility that positive selective pressure might account for the resequenced CRTRs by calculating Fay and Wu's H statistic for each region (Fay and Wu 2000). H compares the nucleotide diversity estimated from heterozygosity (π) against nucleotide diversity estimated from the allele frequency of the derived (nonancestral) allele at each position (θ_H). Significantly negative values of H indicate an excess of high-frequency–derived alleles, consistent with recent, positive selection. Evaluating the significance of a given H depends upon the number of samples and the number of polymorphisms analyzed, so for each of the resequenced regions and in each of the three populations, we simulated 10,000 sample data sets under a standard neutral model, with constant population size and no recombination. In five out of six regions (*EDAR*, *GCG*, *CLSPN*, BX115137, and AW183861), significantly negative H statistics ($P = 0.05$) were observed in the appropriate populations, and in the AD CRTR (*CTPS*, *FLJ23878*, and *SCMH1*), the negative H statistic approached significance for both *FLJ23878* and *SCMH1* (Table 3). Taken altogether, the co-occurrence of (1) an unusually low Tajima's D, (2) an unusually high proportion of monomor-

Table 2. Regions selected for targeted resequencing

					Tajima's D			θ^d ($\times 10^{-4}$)			π^e($\times 10^{-4}$)		
Region	Genome coordinates[a]	Pop[b]	Gene	bp[c]	AD	ED	XD	AD	ED	XD	AD	ED	XD
1	Chr1: 35,270,001–36,160,000	ED	*CLSPN*	16308	−1.29	**−2.44**[f]	−1.07	5.25	2.49	2.07	3.26	0.57	1.35
2	Chr1: 41,130,001–41,420,000	AD	*CTPS*	26417	−0.90	0.15	−0.05	7.08	6.06	5.71	5.29	6.30	5.63
2	Chr1: 41,130,001–41,420,000	AD	*FLJ23878*	7266	−1.66	0.62	−0.11	7.75	7.13	5.89	3.85	8.48	5.69
2	Chr1: 41,130,001–41,420,000	AD	*SCMH1*	30057	**−2.00**	−0.49	−0.85	7.87	6.30	5.92	3.47	5.44	4.51
3	Chr2: 108,400,001–109,070,000	XD	*EDAR*	22094	−0.46	**−2.06**	**−2.40**	13.26	9.48	3.06	11.58	4.02	0.87
4	Chr2: 162,870,001–163,190,000	ED	*GCG*	13293	−0.46	−1.77	−0.96	6.27	4.07	5.59	5.43	1.88	4.00
5	Chr11: 37,820,001–38,290,000	XD	*BX115137*	6422	−0.01	0.77	**−2.60**	13.68	9.12	7.37	13.65	11.24	1.43
6	Chr4: 32,930,000–33,560,000	ED	*AW183861*	21630	−1.28	−1.92	**−2.07**	5.73	2.60	1.88	3.63	1.08	0.65

[a]May 2004 build, hg17.
[b]Population with observed CRTR.
[c]Base pairs resequenced.
[d]Watterson's nucleotide diversity per base pair estimated from the number of segregating sites.
[e]Kimura's nucleotide diversity per base pair estimated from the average heterozygosity per site.
[f]Values of Tajima's D less than -2, the theoretical 95% lower bound for Tajima's D are shown in bold.

Table 3. Fay and Wu's H in resequenced regions

	AD			ED			XD		
Gene	Sn[a]	H[b]	P[c]	Sn[a]	H[b]	P[c]	Sn[a]	H[b]	P[c]
CLSPN	31	2.732	0.4541	14	−3.065	**0.016**[d]	14	−0.738	0.085
CTPS	67	1.825	0.1981	54	2.773	0.2676	52	2.106	0.2332
FLJ23878	24	1.843	0.0676	21	0.434	0.1885	17	−0.814	0.0935
SCMH1	82	5.178	0.0674	67	1.199	0.1758	63	−0.692	0.125
EDAR	109	0.537	0.136	78	−30.213	**0.0055**	27	−21.583	**0.0000**
GCG	30	1.506	0.0833	18	−6.457	**0.0041**	28	−3.741	**0.0347**
BX115137	35	3.456	0.5225	23	1.544	0.3535	20	−7.013	**0.0047**
AW183861	48	0.324	0.1557	23	−7.547	**0.0055**	16	−0.016	0.1514

[a]Number of polymorphic sites identified in resequencing data for which the ancestral allele could be unambiguously identified in chimp.
[b]Fay and Wu's H for sites with ancestral allele data.
[c]Proportion of simulations with more negative Fay and Wu's H than observed in 10,000 simulations.
[d]*P*-values below 0.05 are shown in bold.

phic sites in the Perlegen data (Table 2), (3) reduced absolute nucleotide diversity, and (4) the significantly negative Fay and Wu's H values suggests that selective pressure probably accounts for most of the CRTRs.

Examination of the resequenced region from *CLSPN* (Fig. 4) revealed a coding SNP (Ser525Asn, position 10710 in Fig. 4B, dbSNP rs7537203) with extremely low serine allele frequency in the ED (4% frequency) population and extremely high serine allele frequency in the XD population (83%). This coding polymorphism was also typed in phase I of the HapMap, with similarly extreme differences in allele frequency between Asian and European populations. Close inspection of the CRTR spanning *CLSPN* in the ED population identified an extreme recombination hotspot conserved between all three populations at the telomeric end of the CRTR (Fig. 5), with a relative recombination rate >1000-fold above background inferred by LDhat (McVean et al. 2004). The centromeric end of the CRTR spans a larger region with modestly elevated recombination. This corresponds with the observed patterns of Tajima's D: At the telomeric boundary, normal diversity returns to average levels immediately, while at the centromeric boundary a more gradual recovery is observed.

In contrast to the *CLSPN* CRTR, complete resequencing of the coding regions from *CTP synthase* (*CTPS*), *FLJ23878* (a predicted gene with supporting spliced EST evidence), and *Sex Comb on midleg homolog 1* (*SCMH1*) failed to identify a single coding variant that was significantly enriched in the AD population. Also in contrast to the *CLSPN* CRTR, very little recombination was observed across the 300 kilobases containing these three genes in any of the three populations ($|D'| = 1$ in >95% of pairwise comparisons), so this region is likely to be a recombination coldspot. However, the small amount of observed recombination in the AD population revealed a clear trend across the CRTR, with the lowest Tajima's D region spanning *SCMH1*. Thus, if a selectively advantageous allele exists in this CRTR, it may lie within the *SCMH1* transcript region, but it does not appear to be a coding polymorphism.

EDAR was selected for resequencing because it showed reduced diversity in the XD population, as well as strikingly high levels of F_{st} between ED and XD populations at a few SNPs. This was confirmed by resequencing, where four SNPs were observed with an allele frequency difference of >85% between populations (SNPs 173, 1663, 2531, 93981; details available at the SeattleSNPs Web site), and three other SNPs showed a difference >50% in allele frequency (SNPs 429, 1158, and 62347). Among the four SNPs with the largest allele frequency differences, the 93891 SNP changes an amino acid, but the change is quite conservative (Val370Ala): It substitutes one small, nonpolar amino acid for another and is predicted to have "benign" effects by Polyphen (Sunyaev et al. 2001). The other three SNPs with the largest allele frequency difference fell within 2000 bp of the transcription start site, suggesting a possible regulatory function for one or more of these SNPs. Although no clear CRTR was detected in this region for the ED population, the nucleotide diversity at *EDAR* was also significantly reduced in ED. This suggests that a CRTR smaller than our definition (<300 kb across) may exist in the ED population, consistent with a weaker selective force or an older event in the ED population.

The majority of identified CRTRs spanned more than one known gene, although 20% of the CRTRs (11 of 55 CRTRs) did not contain a "Known Gene" in the UCSC Genome Browser track. For example, the CRTR on chromosome 11 in the XD population (chr11, 37820000–38290000) did not contain any known genes or RefSeq entries and had only one spliced EST with multiple exemplars (GenBank BX115137 at chr11, 37,916,727–37,932,789). Resequencing the region spanning BX115137 confirmed the significantly low diversity in this region (Tajima's $D = -2.60$ in the XD population) (Table 2), thereby confirming findings from the Perlegen data set. This CRTR might reflect direct selective pressure upon this EST, or it might reflect selection upon a long-range regulatory element affecting expression of genes outside of the CRTR in the XD population. It is worth noting that the two of the closest genes to this CRTR are *RAG1* and *RAG2* (chr11, 36,546,150–36,557,871, and chr11, 36,570,070–36,576,362, respectively), which are essential for adaptive immunity through the rearrangement of *T Cell Receptor* genes (Fugmann 2001). Therefore, it is possible that long-range regulatory elements affecting the function of these genes could be subject to strong selective pressure.

Discussion

Identifying regions of the human genome that have experienced substantial selective pressure can provide insights into the location of functionally important polymorphisms and may help prioritize targets for association mapping (Sabeti et al. 2002; Clark et al. 2003b). This is especially true in genomic regions where selective pressure has been experienced in a geographically restricted manner, where large allele frequency differences in functional variation can exist between geographic subpopulations (Akey et al. 2004; Stajich and Hahn 2005). We demonstrated a

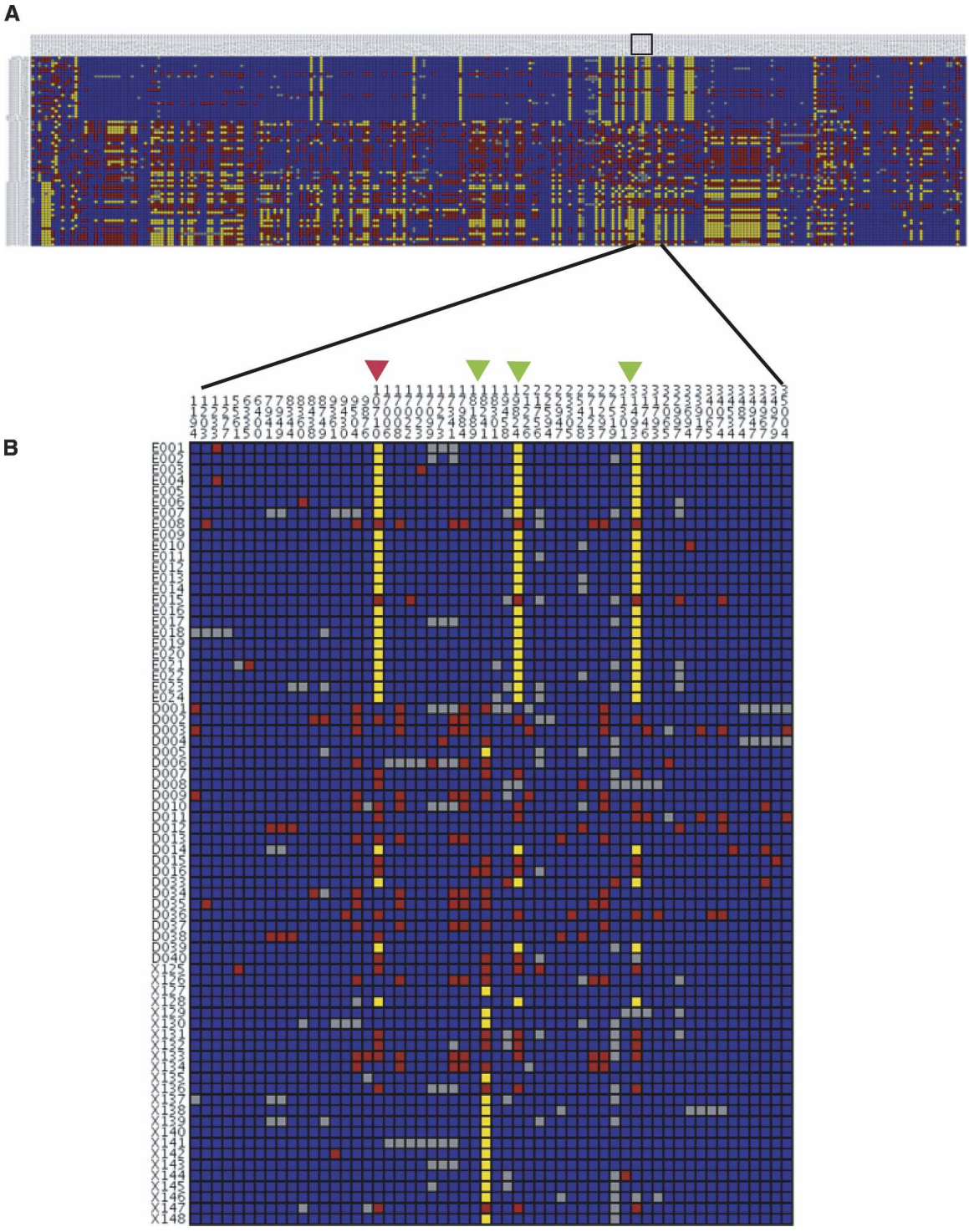

Figure 4. (*A*) A visual genotype for 1.5 Mbp spanning the *CLSPN* CRTR in the Perlegen data. Each row corresponds to an individual, and each column corresponds to a polymorphic site, with genotypes color coded as follows: Common allele homozygotes are shown in blue, heterozygotes are shown in red, rare allele homozygotes are shown in yellow, and missing data are shown as gray. The *top* 24 samples are ED, the *middle* 23 samples are AD, and the *bottom* 24 samples are XD. Although nucleotide diversity is depressed across a large region, there is no clear minimum within the CRTR. Nucleotide diversity was relatively constant across the region, so *CLSPN* (shown as a black box) was selected as a target for resequencing because of interesting patterns of F_{st} between ED and XD, in addition to low nucleotide diversity. (*B*) A visual genotype of the resequencing results for the *CLSPN* gene. The *top* 24 samples are ED; the *middle* 24 samples are AD, and the *bottom* 24 samples are XD. As expected, a number of polymorphisms nearly fixated between ED and XD were observed. One of these SNPs (10710, red arrowhead) changes an amino acid (Ser525Asn), whereas the other three are intronic (green arrowheads).

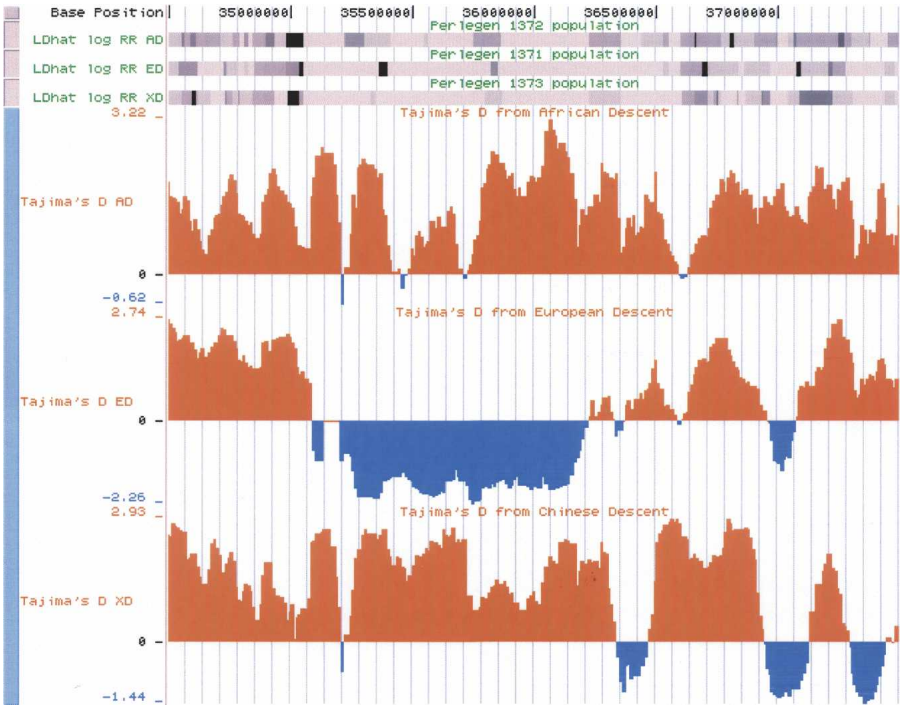

Figure 5. A close-up of the *CLSPN* CRTR from the UCSC genome browser is shown, with the Tajima's D tracks as well as a set of tracks showing the inferred relative recombination rate from LDhat for each population in grayscale (track label, LDhat log RR AD/ED/XD): Darker segments correspond to high inferred recombination rates. *CLSPN* is located at 35.9 Mbp. The *left* edge of the *CLSPN* CRTR (at ~35 Mbp in the ED population) corresponds to a strong recombination hotspot observed in all three populations, but of greater interest are the hotspots spanned by the CRTR at ~35.4 Mbp and ~35.8 Mbp. Thus, although this CRTR does span a region with reduced recombination overall, there are several inferred hotspots within the CRTR that are shared between populations.

significant correlation between Tajima's D in full resequencing data from SeattleSNPs and dense genotyping data from Perlegen. This correlation was exploited to identify large regions of the genome where strong selective pressures have recently been experienced by at least one of the three populations surveyed: AD, ED, and XD. A 100-kbp sliding window analysis was used to identify CRTRs in each population, representing large regions with SFS enriched for low-frequency polymorphisms. We selected eight candidate genes or ESTs within a subset of these CRTRs for resequencing analysis, and although the parameters used to define a CRTR were not theoretically derived, they appear to be conservative in detecting positive selective pressure, given the resequencing results. Thus, we demonstrate that SFSs from dense genotyping data are a useful way to screen the genome for regions that have probably experienced recent and strong selective pressure.

The observed distributions of the windowed Tajima's D values were remarkably similar in the ED and XD populations (Fig. 3), suggesting not only that the demographic histories of these populations are similar, as has previously been suggested (Yu et al. 2001; Marth et al. 2004), but that the degree of selective pressure on these populations may also have been similar. However, the precise regions of the genome that fell into the tails of the distribution were generally not the same between these two populations. Thus, to the extent that the distribution of Tajima's D was affected by selective pressure in these populations, the selective pressures appear to have acted on independent regions in each population, with a few exceptions, including the three ED/XD shared CRTRs identified in Table 1. In contrast to findings in the ED and XD populations, the mean Tajima's D distribution was lower, and the range of this distribution was substantially narrower in the AD population, suggesting a unique demographic and/or selective history for this population. Closer inspection of the AD distribution demonstrates less correlation between adjacent regions, as well as a more restricted range of Tajima's D, consistent with reports of shorter-range LD in this population attributable to a larger effective population size (Harpending and Rogers 2000; Carlson et al. 2003; Marth et al. 2004).

Data from these three major populations showed significant differences in the quantity of CRTRs between populations. Overall, relatively few CRTRs were observed in the AD population, and the observed CRTRs were generally smaller than were those in the ED and XD populations. This could be due to less dramatic selective pressure on African populations in recent evolutionary history, but we consider it unlikely that selective pressure from pathogens or diet is substantially weaker in this population. If anything, the pathogen load might be expected to be highest in regions where humans have lived the longest, although the degree of mortality from pathogens may have been attenuated. Alternatively, admixed European chromosomes might serve to obscure Africa-specific selective sweeps, but this possibility also seems unlikely because F_{st} is low and few if any polymorphisms have fixated between these populations. Thus, admixture between European and African populations tends to reduce Tajima's D by increasing the number of relatively rare SNPs, and admixture from the European population should actually have enhanced detection of AD-specific CRTRs. Demographic parameters such as a population bottleneck in the Eurasian populations (Marth et al. 2004) may also have enhanced detection of selective events in the non-African populations. Also, the higher average Tajima's D in ED populations may simply facilitate the identification of CRTRs in ED, relative to the AD population. A final possibility is that the relatively larger effective population size of the AD population allows greater opportunity for recombination between a selectively advantageous allele and neighboring regions during the course of a selective sweep. In a larger population, the ancestral haplotype bearing the advantageous allele is exposed to recombination in a larger number of individuals, and therefore, the segment of ancestral haplotype that eventually fixates will be shorter than in a smaller population. This hypothesis is supported by the observation that several genes with reportedly significant negative Tajima's D values in the AD population (*FY* [Hamblin and Di Rienzo 2000; Hamblin et al. 2002] and *APOA2* [Fullerton et al. 2002]) were not observed within CRTRs in the AD population at a 1% empiric threshold, or at an even less stringent 5% threshold. Taken together with the relative infrequency of CRTRs in the AD

population, this suggests that CRTRs in the AD population may cover shorter segments of the genome than in the ED and XD populations, and will therefore require smaller windows for detection. As the density of genotyping data sets increases (e.g., phase 2 of the HapMap will type millions of additional SNPs, to add to the more than one million SNPs typed in phase 1) (Gibbs et al. 2003), the sliding window size of this type of analysis can be reduced, and smaller CRTRs may become detectable in African populations.

Genes within the CRTR regions may provide important targets for genotype/phenotype studies. For example, *CYP3A4* and *CYP3A5* play a central role in the metabolism of some prescribed drugs, lie within a CRTR in the ED population, and have been shown to have significantly low Tajima's D in European samples (Thompson et al. 2004). Another example is *VKORC1*, which has been linked to human warfarin dosing and shown to have low haplotype diversity in Asian populations (Rieder et al. 2005), and is located in a CRTR on chromosome 16 in the XD population. Although we report CRTRs at a 1% empiric threshold, several previously reported genes with significantly negative Tajima's D in European populations could be identified in CRTRs at a less stringent 5% empiric threshold. For example, lactase (LCT) lies within a ~1 Mbp CRTR in the ED population at the 5% threshold and has been suggested as a target of selective pressure in the European population (Bersaglieri et al. 2004). Furthermore, a cluster of genes previously suggested as subject to natural selection in Europeans (*KEL*, *TRPV5*, *TRPV6*, and *EPHB6*) (Akey et al. 2004; Stajich and Hahn 2005) exhibited low Tajima's D in the ED population, but not across a large enough segment to meet our definition of a CRTR. Based on this combined evidence, CRTRs observed at the more stringent 1% empiric threshold (Table 1) seem quite likely to represent selective sweeps.

Resequencing of candidate genes within a number of the CRTRs provides further support for this hypothesis: In every CRTR selected for resequencing, at least one gene with a dramatic departure from the expected distribution of Tajima's D under neutrality was observed (Table 2), and in most CRTRs, a significant departure from the expected distribution of Fay and Wu's H was also observed (Table 3). In addition to theoretical evidence, the genes resequenced in the CRTRs also fell within the bottom 5% of the empirical distribution of Tajima's D in >170 genes resequenced by SeattleSNPs (Fig. 1). For example, the resequencing-based Tajima's D for *SCMH1* was the lowest Tajima's D value that we have ever observed in the AD population, compared with data from 179 resequenced genes. Furthermore, the Tajima's D for genes selected from ED CRTRs was similar to the Tajima's D for genes previously demonstrated to be robustly incompatible with neutrality in several studies (Akey et al. 2004; Stajich and Hahn 2005).

Within each CRTR, it is apparent that a single common haplotype has recently increased dramatically in frequency, at the expense of all other haplotypes within the CRTR. However, it is not yet clear whether the fitness advantage is attributable to a genotype at a single SNP or a haplotype of multiple SNPs. Haplotype effects are more plausible when a single transcript spans the majority of a CRTR (e.g., *PHKB* on chromosome 16 in XD) or across regions containing groups of functionally related genes (e.g., the Olfactory Receptor gene cluster contained within the AD CRTR on chromosome 11). However, the majority of CRTRs contain multiple genes without clearly related gene functions, so we have not pursued analyses of gene function within the CRTRs, because it is likely that many if not most of the genes within the CRTRs simply represent hitchhiking events where the advantageous allele within a single gene swept the neighboring genes along with it as it increased in frequency (Fay and Wu 2000). Identification of viable candidates for these selective effects would at a minimum require complete resequencing of all functional regions within a CRTR and follow-up of any potentially interesting variants. As more complete resequencing data become available within each CRTR, other SFS test statistics, such as the Fay and Wu's H test, can potentially be applied to each CRTR to narrow the candidate interval containing the advantageous variant.

The pattern of concordance for CRTRs between ED and XD populations was also interesting: An overlap of three CRTRs between these populations is quite striking, given that the CRTRs comprise <1% of the genome. Given that Tajima's D values from the resequencing data are in the range of genes previously reported to be inconsistent with neutrality under a range of demographic parameters, we believe that the shared CRTRs probably represent shared selective pressures between these populations. However, not all shared CRTRs necessarily represent a single selective sweep in multiple populations. For example, Tajima's D was significantly low at *EDAR* in both XD and ED populations, but the genomic extent of the CRTR is substantially greater in XD than ED populations. This could represent sweeps that occurred at different times historically, but the extreme F_{st} at a series of polymorphisms in *EDAR* is consistent with either a divergent sweep with one haplotype favored in ED and a different haplotype favored in XD, or parallel sweeps favoring an allele shared by both haplotypes but not by other African haplotypes (e.g., site 96563 in the 3′ UTR, rs1478517). No CRTRs were observed to be shared between all three populations, but this would be consistent with the ascertainment bias toward high-frequency SNPs: If no high-frequency SNPs exist in any of the three populations, then no SNPs were available for use in the Perlegen data. Therefore, although global CRTRs were not observed, such regions may be present as large regions without genotype data in the Perlegen data set.

Considering each of the regions resequenced, identification of the specific SNP or SNPs conferring a selective advantage is not trivial. For example, although the patterns of SFS suggest that *SCMH1* is likely to harbor the advantageous allele responsible for the selective sweep that created the CRTR in the AD population, no coding polymorphism was identified in the AD population with significantly enriched allele frequency in either *SCMH1* or the other two genes. Given that we resequenced all of the known coding regions in this CRTR, if a polymorphism within *SCMH1* drove the sweep, then the function of the polymorphism was probably regulatory rather than structural. In contrast, the *CLSPN* and *EDAR* resequencing data identified interesting candidate cSNPs, and selective pressure on these cSNPs could conceivably account for the *EDAR* and *CLSPN* CRTRs. More extensive resequencing within each CRTR is required to determine whether other candidate SNPs exist in neighboring genes.

In conclusion, the availability of adequately dense genotyping data sets clearly facilitates the identification of regions of the human genome with unusual SFS, which may have been subjected to strong positive selective pressure in the recent past. Current data appear to be adequate to identify such regions in ED and CD populations, but denser data will be necessary for analysis of AD populations, probably due to the larger effective population size of this population. Detection of regions subject to balancing selection (e.g., *HLA*) or with less complete selective

sweeps (e.g., *FY*) will probably require a substantially denser data set than is currently available. Although most CRTRs span multiple genes, within each CRTR the selective sweep favored only one haplotype at the expense of all others, so a single selectively advantageous polymorphism in a single gene could conceivably account for each CRTR, with the reduced diversity in flanking regions representing a hitchhiking event. Dissection of the underlying functional variant (or variants) within each CRTR may require comprehensive resequencing within the CRTR to identify candidate functional variation, but where it is feasible, functional analysis of a priori functional variants (e.g., the *CLSPN* Ser525Asn cSNP) should substantially accelerate this process.

Methods

Samples

Twenty-four individuals from each of three populations were resequenced: 24 African American individuals from the Coriell HD100AA diversity panel (population AD), 24 CEPH individuals (population ED), and 24 Chinese Americans from the Coriell HD100A diversity panel (population XD). All ED individuals overlap with the Perlegen European panel (dbSNP population 1371); all but one of the AD individuals overlap with the Perlegen African American panel (dbSNP population 1372); and all XD individuals overlap with the Perlegen Chinese panel (dbSNP population 1373). Coriell accession numbers are as follows: AD population, NA17101–NA17116, NA17133–NA17140; ED population, NA06990, NA07019, NA07348, NA07349, NA10830, NA10831, NA10842, NA10843, NA10844, NA10845, NA10848, NA10850, NA10851, NA10852, NA10853, NA10854, NA10857, NA10858, NA10860, NA10861, NA12547, NA12548, NA12560, NA17201; and XD population, NA17733–NA17747, NA17749, NA17752–NA17757, NA17759, NA17761.

Perlegen data

All genotype data for populations 1371 (ED), 1372 (AD), and 1373 (XD) from build 124 of dbSNP were downloaded and parsed for analysis. Several quality controls were applied to the data set to verify that the data were completely and accurately retrieved. Comparison of the 46 samples overlapping between Perlegen and SeattleSNPs identified 95,354 concordant genotypes out of 96,170 genotypes compared (>99.2%; consistent with Hinds et al. 2005). We also observed good concordance between the total number of Perlegen SNPs retrieved from dbSNP (1.58 million) and the numbers reported by Hinds et al. The Perlegen SNP class data for each SNP were downloaded for each autosome from the Perlegen Web site (http://genome.perlegen.com/browser/download.html).

Sequencing analysis

Methods for sequence analysis of candidate genes obtained from SeattleSNPs (http://pga.gs.washington.edu) have previously been described in detail (Carlson et al. 2003, 2004; Livingston et al. 2004). Briefly, PCR amplicons were tiled across the full genomic sequence of a gene or selected genomic regions and sequenced by using standard BigDye Terminator v3.1 methods. The amplification primers and sequencing protocol are available at http://pga.gs.washington.edu. Sequence data from the ABI 3730XL instrument were base called with Phred (Ewing et al. 1998) and assembled onto a reference sequence with the add reads feature in Consed (Gordon et al. 1998). Polymorphisms were identified by PolyPhred v4.29 (Nickerson et al. 1997), and Consed was used to visualize the sequences and confirm polymorphisms.

Nucleotide diversity analysis

There are several statistics that can be used to describe nucleotide diversity, including θ_s (equation 1), π (equation 2), and θ_H (equation 3). These statistics can be calculated for a given resequencing data set by using the following parameters: n is the number of chromosomes resequenced, Sn is the number of polymorphic sites observed, p_i is the derived (nonancestral) allele frequency of the ith SNP, and q_i is the ancestral allele frequency of the ith SNP.

$$\theta_s = \frac{Sn}{\sum_{i=1}^{n-1}\frac{1}{i}} \tag{1}$$

$$\pi = \frac{n}{n-1}\sum_{i=1}^{Sn} 2p_i q_i \tag{2}$$

$$\theta_H = \frac{n}{n-1}\sum_{i=1}^{Sn} 2p_i^2 \tag{3}$$

There are many statistics that can evaluate departures from the expected patterns of neutral variation. One of these is Tajima's D (Tajima 1989), equation 4:

$$D = \frac{\pi - \theta_s}{\sqrt{Var(\pi - \theta_s)}} \tag{4}$$

The theoretical distribution of Tajima's D (95% confidence interval between −2 and +2) assumes that polymorphism ascertainment is independent of allele frequency. High values of Tajima's D suggest an excess of common variation in a region, which can be consistent with balancing selection or population contraction. Negative values of Tajima's D, on the other hand, indicate an excess of rare variation, consistent with population growth, or positive selection. Population admixture can lead to either high or low Tajima's D values in theory. Demographic parameters would be expected to affect the genome more evenly than selective pressures, so previous analyses have suggested that using the empiric distribution of Tajima's D from a collection of regions across the genome provides advantages in assessing whether selection or demography might explain an observed deviation from expectation (Akey et al. 2004; Stajich and Hahn 2005). Because of the ascertainment bias toward common polymorphism in the Perlegen data set, extremely positive Tajima's D values are difficult to interpret, and modeling ascertainment is difficult. However, given that the ascertainment bias raises the mean of the distribution, extreme negative values in extended regions can be useful in qualitatively identifying interesting regions for full resequencing and more rigorous theoretical analysis of nucleotide diversity.

For genic comparisons between SeattleSNPs data and Perlegen data, Tajima's D in the SeattleSNPs data was calculated for all observed polymorphic sites. The median transcript size in the SeattleSNPs data set is 14,649 bp, and on average, an additional 3000 bp of flanking sequence was also resequenced, for a median analyzed region of 17.5 kbp. Given the density of the Perlegen map, the median number of polymorphic SNPs in the resequenced region was only six in the ED population and seven in the AD population, a rather small number of polymorphisms for Tajima's D estimation. One hundred nineteen out of 179 genes analyzed in the AD population had five or more polymorphic Perlegen SNPs within the resequenced region, and in the ED population, 107 genes had five or more polymorphic Perlegen

SNPs. Significant linkage disequilibrium routinely extends 10 kb, even in the AD population, so patterns of nucleotide diversity should be conserved over similar distances. We extended the region analyzed from the Perlegen data to include 10 kb upstream and 10 kb downstream of the transcript, under the expectation that this would increase the number of sites per gene and therefore the accuracy of the Tajima's D estimate. Expanding the region raised the median number of polymorphic sites per gene to 16 in ED and 19 in AD. As expected, the larger number of sites per gene improved the correlation between the SeattleSNPs and Perlegen data substantially. In the AD population, $R^2 = 0.29$ in extended versus $R^2 = 0.04$ in the transcript for the 119 genes with five or more polymorphic SNPs in the transcript, and in the ED population, $R^2 = 0.66$ in extended versus $R^2 = 0.15$ in transcript for the 107 genes with five or more polymorphic SNPs in the transcript. Thus, for the genic comparison in Figure 1, Tajima's D in the Perlegen data was calculated on the basis of all observed polymorphic sites within 10 kb of the longest reported transcript in Entrez Gene. Within each population, only genes with five or more polymorphic SNPs in the Perlegen data were included in the comparison of data sets, which yielded 178 genes in the AD population and 173 genes in the ED population.

For the windowed analysis of the genome, Tajima's D was calculated independently in each population. Sliding windows of 100 kb were analyzed across all autosomal regions in the Perlegen data, stepping by 10 kb. Thus, the first window evaluated on chromosome 1 was genome coordinates chr1, 1–100,000; the second window was genome coordinates chr1, 10,001–110,000; and so forth. Because adjacent windows overlap, in the genome browser track, the Tajima's D is reported for each window using the coordinates of the central 10 kb. Thus, the observed Tajima's D for window chr1, 1–100,000, is reported at chr1, 45,001–55,000. These data have been made available as a track in the UCSC genome browser (http://genome.ucsc.edu/).

The empirically determined distribution of Tajima's D within the sliding windows was used to identify CRTRs, defined as a region of ≥20 contiguous windows where >75% of the windows were in the bottom 1% of the empirical Tajima's distribution. The empirically determined bottom 1% of Tajima's D values corresponded to Tajima's $D < -0.70$ in the AD population, Tajima's $D < -1.49$ in the ED population, and Tajima's $D < -1.743$ in the XD population. Window size was chosen to provide a reasonably large number of SNPs per window (average 54.3 SNPs per window in AD, 47.1 in ED, and 42.8 in XD). Under the restriction that 75% of all windows must be below threshold within a contiguous region, the distribution of contiguous region sizes for each population is shown in Supplemental Figure 1: In each population, <10% of all such contiguous regions exceeded 20 windows in length, so we used a threshold of 20 adjacent windows to define a CRTR. Given that each window overlaps its neighbor by 90,000 bp, a stretch of 20 contiguous windows corresponds to 300,000 bp.

Fay and Wu's H analysis

The chimpanzee allele was used to determine the ancestral human allele within the eight resequenced regions. Only SNPs where the chimpanzee alignment was unambiguous and matched one of the existing human alleles were used in this analysis. In each population, 24 samples were resequenced, representing 48 chromosomes, so Fay and Wu's H ($H = \pi - \theta_H$) was calculated within each population and then compared against simulated data. Simulations were run for each region in each of the three populations, under the conservative assumption of no recombination. Simulations were performed by using mksamples (Hudson 2002), assuming 48 chromosomes and the appropriate number of SNPs (*Sn*) with ancestral allele information (Table 3). For each analysis, 10,000 simulations were performed, and the proportion of simulations with more negative H than the observed data (equivalent to a *P*-value) is shown in Table 3.

Recombination rate analysis

Recombination was addressed in two ways: Large-scale recombination rates for the region spanning each CRTR were estimated from the deCODE sex averaged recombination rate (Kong et al. 2002) for the flanking megabase, when the CRTR fell within a single deCODE interval, and by weighted averages of the flanking megabase when the CRTR spanned a boundary between deCODE intervals. At a much finer scale, the relative recombination rate per nucleotide was estimated across 3 Mb flanking the *CLSPN* CRTR by using the pairwise option within the LDhat program (McVean et al. 2004). This program first estimates the coalescent likelihood of observing each pair of segregating sites, treating each pair as independent, and then estimates the recombination rate for the entire region over a grid. Polymorphism data from chr1, 34,500,000–37,500,000, was analyzed independently for each population. The estimated fine-scale recombination rates across this region for each population are shown in Figure 5.

Acknowledgments

This work was supported by a Program for Genomic Applications grant from the National Heart, Lung, and Blood Institute (HL66682 and HL66642 to D.N. and M.R.). D.T. was supported by grants from the National Human Genome Research Institute (IP41HG02371 and HG02238 to David Haussler). We thank Dana Crawford, Alex Reiner, and Eric Torskey for comments on the manuscript, as well as the entire SeattleSNPs resequencing team for their extraordinary efforts on this project.

References

Akey, J.M., Eberle, M.A., Rieder, M.J., Carlson, C.S., Shriver, M.D., Nickerson, D.A., and Kruglyak, L. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* **2:** e286.

Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E., and Hirschhorn, J.N. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74:** 1111–1120.

Carlson, C.S., Eberle, M.A., Rieder, M.J., Smith, J.D., Kruglyak, L., and Nickerson, D.A. 2003. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat. Genet.* **33:** 518–521.

Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L., and Nickerson, D.A. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74:** 106–120.

Clark, A.G., Glanowski, S., Nielsen, R., Thomas, P., Kejariwal, A., Todd, M.J., Tanenbaum, D.M., Civello, D., Lu, F., Murphy, B., et al. 2003a. Positive selection in the human genome inferred from human–chimp–mouse orthologous gene alignments. *Cold Spring Harb. Symp. Quant. Biol.* **68:** 471–477.

———. 2003b. Inferring nonneutral evolution from human–chimp–mouse orthologous gene trios. *Science* **302:** 1960–1963.

Ewens, W.J. 1972. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3:** 87–112.

———. 1979. *Mathematical population genetics*. Springer-Verlag, New York.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred, I: Accuracy assessment. *Genome Res.* **8:** 175–185.

Fay, J.C. and Wu, C.I. 2000. Hitchhiking under positive Darwinian selection. *Genetics* **155:** 1405–1413.
Fu, Y.X. and Li, W.H. 1993. Statistical tests of neutrality of mutations. *Genetics* **133:** 693–709.
Fugmann, S.D. 2001. RAG1 and RAG2 in V(D)J recombination and transposition. *Immunol. Res.* **23:** 23–39.
Fullerton, S.M., Clark, A.G., Weiss, K.M., Taylor, S.L., Stengard, J.H., Salomaa, V., Boerwinkle, E., and Nickerson, D.A. 2002. Sequence polymorphism at the human apolipoprotein AII gene (APOA2): Unexpected deficit of variation in an African-American sample. *Hum. Genet.* **111:** 75–87.
Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F., Yang, H., Ch'ang, L.Y., Huang, W., Liu, B., Shen, Y., et al. 2003. The International HapMap Project. *Nature* **426:** 789–796.
Gordon, D., Abajian, C., and Green, P. 1998. Consed: A graphical tool for sequence finishing. *Genome Res.* **8:** 195–202.
Hamblin, M.T. and Di Rienzo, A. 2000. Detection of the signature of natural selection in humans: Evidence from the Duffy blood group locus. *Am. J. Hum. Genet.* **66:** 1669–1679.
Hamblin, M.T., Thompson, E.E., and Di Rienzo, A. 2002. Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.* **70:** 369–383.
Harpending, H. and Rogers, A. 2000. Genetic perspectives on human origins and differentiation. *Annu. Rev. Genomics Hum. Genet.* **1:** 361–385.
Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A., and Cox, D.R. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* **307:** 1072–1079.
Hudson, R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18:** 337–338.
Hughes, A.L. and Yeager, M. 1998. Natural selection and the evolutionary history of major histocompatibility complex loci. *Front. Biosci.* **3:** d509–d516.
Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996–1006.
Kimura, M. 1983. *The neutral theory of molecular evolution.* Cambridge University Press, Cambridge, UK.
Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* **31:** 241–247.
Libert, F., Cochaux, P., Beckman, G., Samson, M., Aksenova, M., Cao, A., Czeizel, A., Claustres, M., de la Rua, C., Ferrari, M., et al. 1998. The deltaccr5 mutation conferring protection against HIV-1 in Caucasian populations has a single and recent origin in Northeastern Europe. *Hum. Mol. Genet.* **7:** 399–406.
Livingston, R.J., von Niederhausern, A., Jegga, A.G., Crawford, D.C., Carlson, C.S., Rieder, M.J., Gowrisankar, S., Aronow, B.J., Weiss, R.B., and Nickerson, D.A. 2004. Pattern of sequence variation across 213 environmental response genes. *Genome Res.* **14:** 1821–1831.
Marth, G.T., Czabarka, E., Murvai, J., and Sherry, S.T. 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166:** 351–372.
McVean, G.A., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R., and Donnelly, P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* **304:** 581–584.
Nickerson, D.A., Tobe, V.O., and Taylor, S.L. 1997. PolyPhred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* **25:** 2745–2751.
Nielsen, R., Hubisz, M.J., and Clark, A.G. 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* **168:** 2373–2382.
Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J., et al. 2005. A Scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS. Biol.* **3:** e170.
Rieder, M.J., Reiner, A.P., Gage, B.F., Nickerson, D.A., Eby, C.S., McLeod, H.L., Blough, D.K., Thummel, K.E., Veenstra, D.L., and Rettie, A.E. 2005. Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. *N. Engl. J. Med.* **352:** 2285–2293.
Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419:** 832–837.
Seltsam, A., Hallensleben, M., Kollmann, A., and Blasczyk, R. 2003. The nature of diversity and diversification at the ABO locus. *Blood* **102:** 3035–3042.
Smith, J.M. and Haigh, J. 1974. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23:** 23–35.
Stajich, J.E. and Hahn, M.W. 2005. Disentangling the effects of demography and selection in human history. *Mol. Biol. Evol.* **22:** 63–73.
Straus, D.S. and Taylor, C.E. 1981. Hitchhiking and linkage disequilibrium between hemoglobin S and nearby restriction sites. *Hum. Hered.* **31:** 348–352.
Sunyaev, S., Ramensky, V., Koch, I., Lathe III, W., Kondrashov, A.S., and Bork, P. 2001. Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10:** 591–597.
Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585–595.
Thompson, E.E., Kuttab-Boulos, H., Witonsky, D., Yang, L., Roe, B.A., and Di Rienzo, A. 2004. CYP3A variation and the evolution of salt-sensitivity variants. *Am. J. Hum. Genet.* **75:** 1059–1069.
Thornton, K. 2005. Recombination and the properties of Tajima's D in the context of approximate likelihood calculation. *Genetics* (in press).
Verrelli, B.C., McDonald, J.H., Argyropoulos, G., Destro-Bisol, G., Froment, A., Drousiotou, A., Lefranc, G., Helal, A.N., Loiselet, J., and Tishkoff, S.A. 2002. Evidence for balancing selection from nucleotide sequence analyses of human G6PD. *Am. J. Hum. Genet.* **71:** 1112–1128.
Watterson, G.A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7:** 256–276.
Weir, B.S. and Cockerham, C.C. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* **38:** 1358–1370.
Williamson, S.H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R., and Bustamante, C.D. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci.* **102:** 7882–7887.
Yu, N., Zhao, Z., Fu, Y.X., Sambuughin, N., Ramsay, M., Jenkins, T., Leskinen, E., Patthy, L., Jorde, L.B., Kuromori, T., et al. 2001. Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. *Mol. Biol. Evol.* **18:** 214–222.

Web site references

http://pga.gs.washington.edu; Seattle SNPs Web site.
http://genome.perlegen.com/browser/download.html; Perlegen Web site.
http://genome.ucsc.edu/cgi-bin/hgGateway; UCSC Genome Browser.

Received June 21, 2005; accepted in revised form September 6, 2005.